A study of the robustness of raw waveform based speaker embeddings under mismatched conditions

Ge Zhu, Frank Cwitkowitz and Zhiyao Duan
University of Rochester

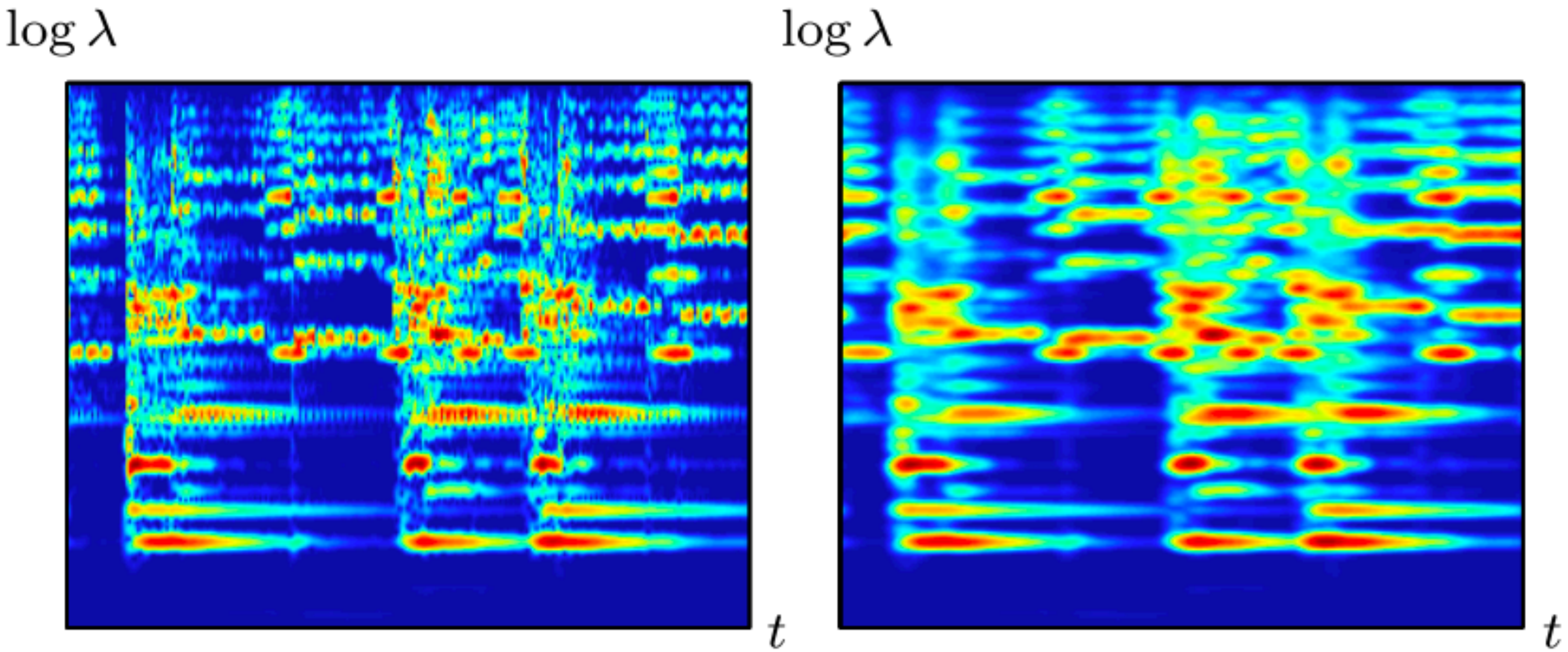
https://github.com/gzhu06/TDspkr-mismatch-study

A study of the robustness of raw waveform based speaker embeddings under mismatched conditions

- Why we are interested in raw waveform?
- Channel mismatch problem
- Proposed strategies
- Experiments

Why we are interested in raw waveform?

- Mel fbank may not optimal:



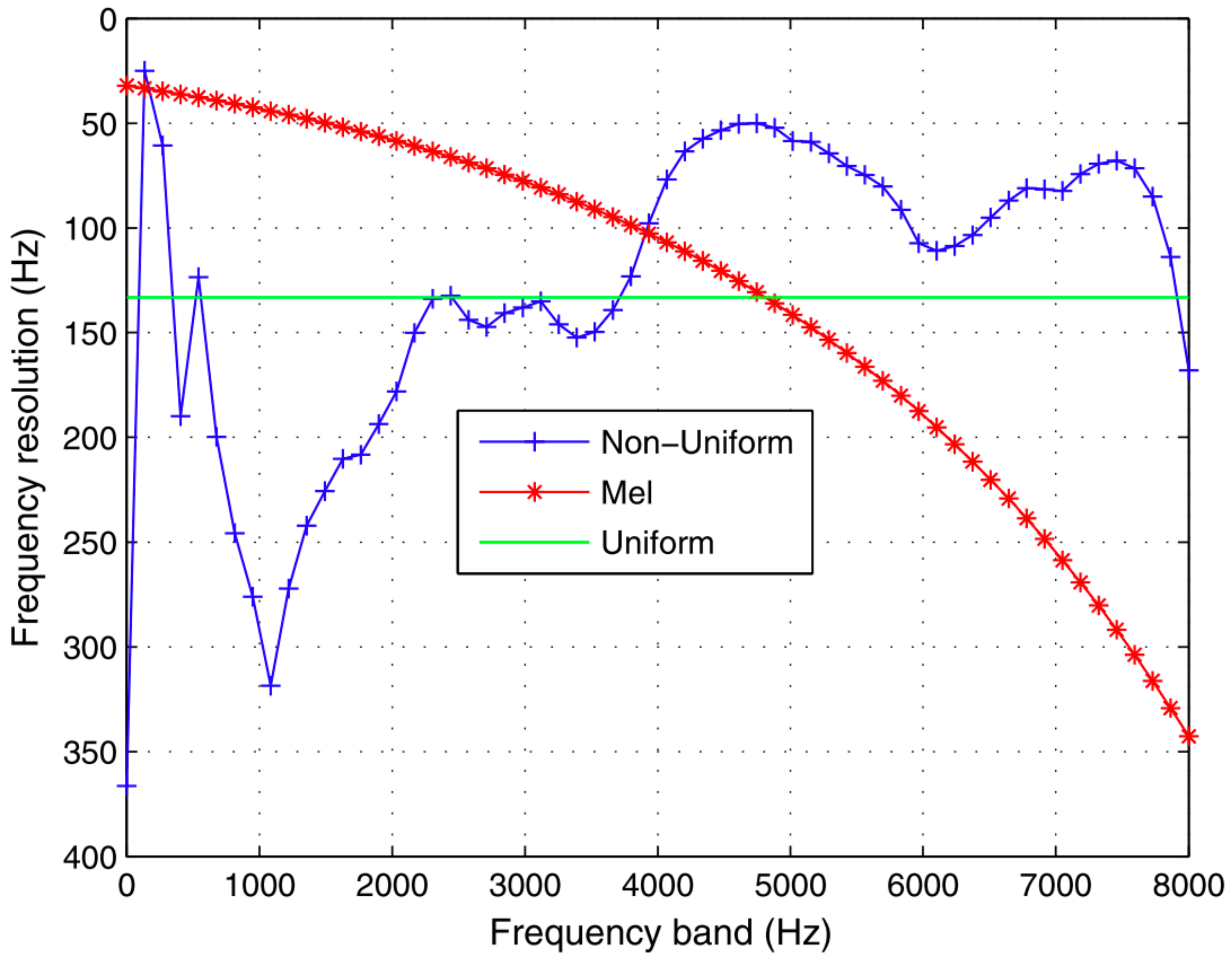
(a) (b)

Scalogram $\log |x \star \psi_\lambda(t)|^2$ Averaged scalogram $\log |x \star \psi_\lambda|^2 \star \phi^2(t)$

Figure: Joakim Andén, Stéphane Mallat. *Deep Scattering Spectrum*. IEEE TRANSACTIONS ON SIGNAL PROCESSING

Why we are interested in raw waveform?

- Frequency resolution in Mel scale



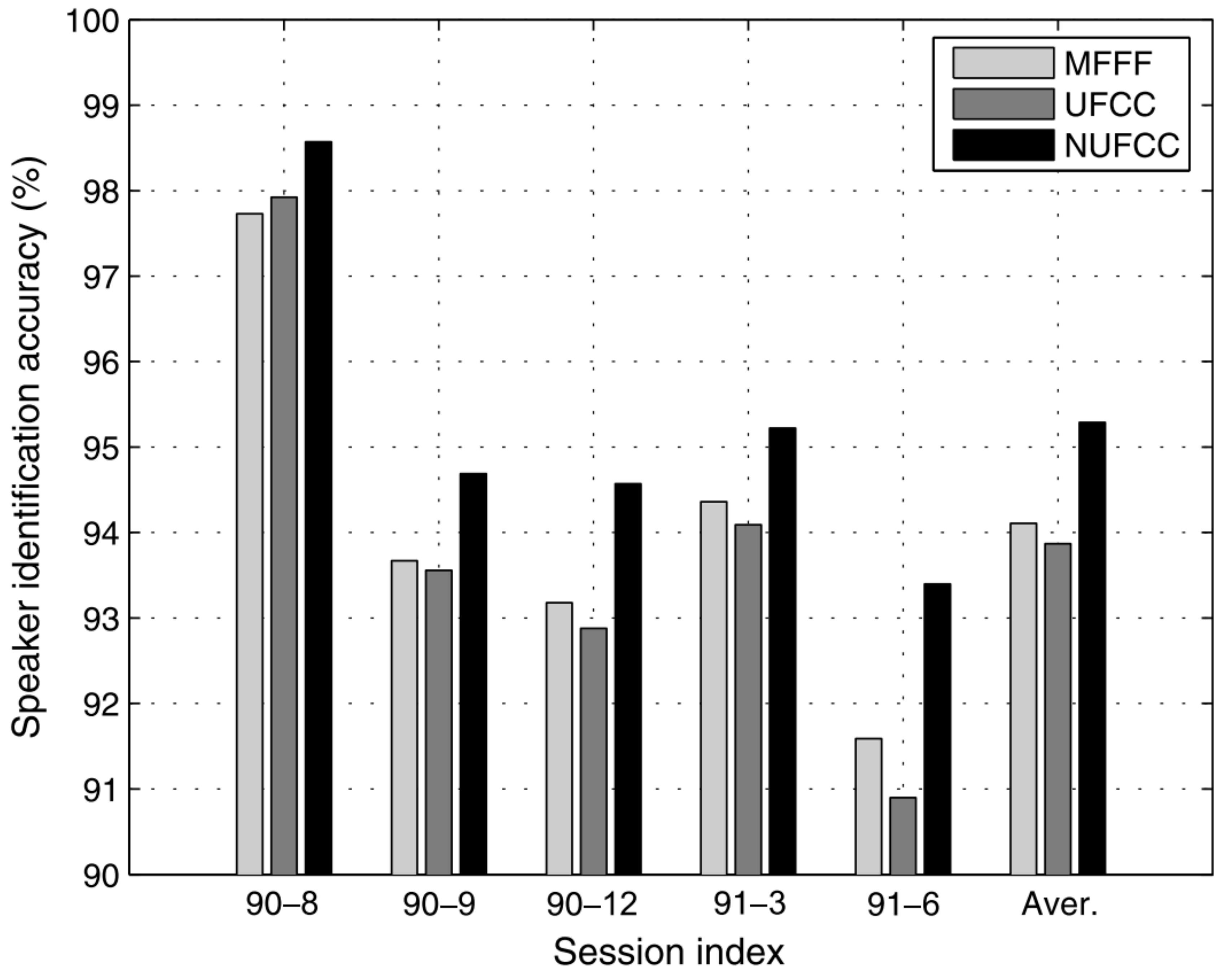


Figure: Xugang Lu, Jianwu Dang. *An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification.* Speech Communication

Why we are interested in raw waveform?

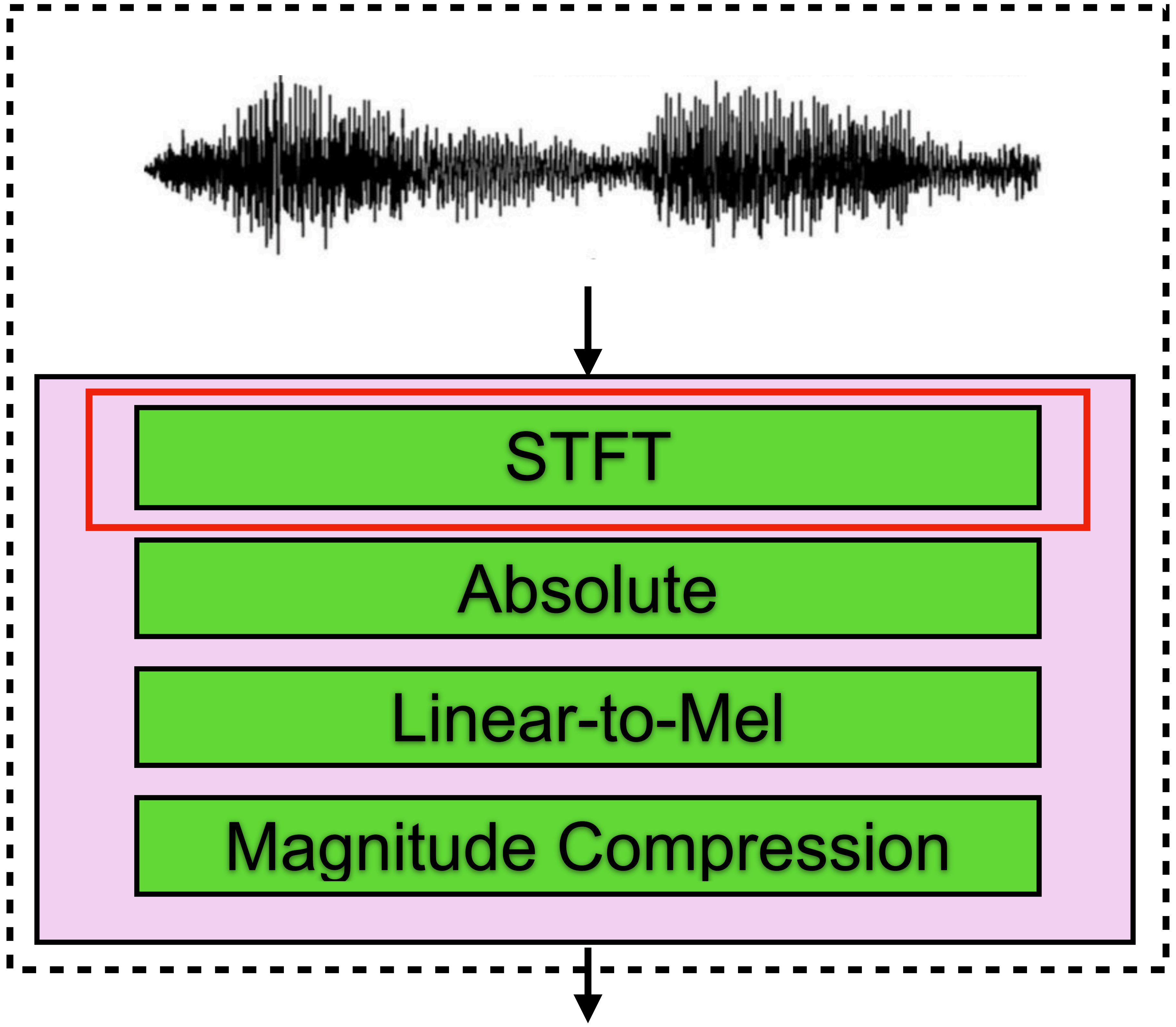
Modern unsupervised/self-supervised speech frontend applies waveform as audio inputs:

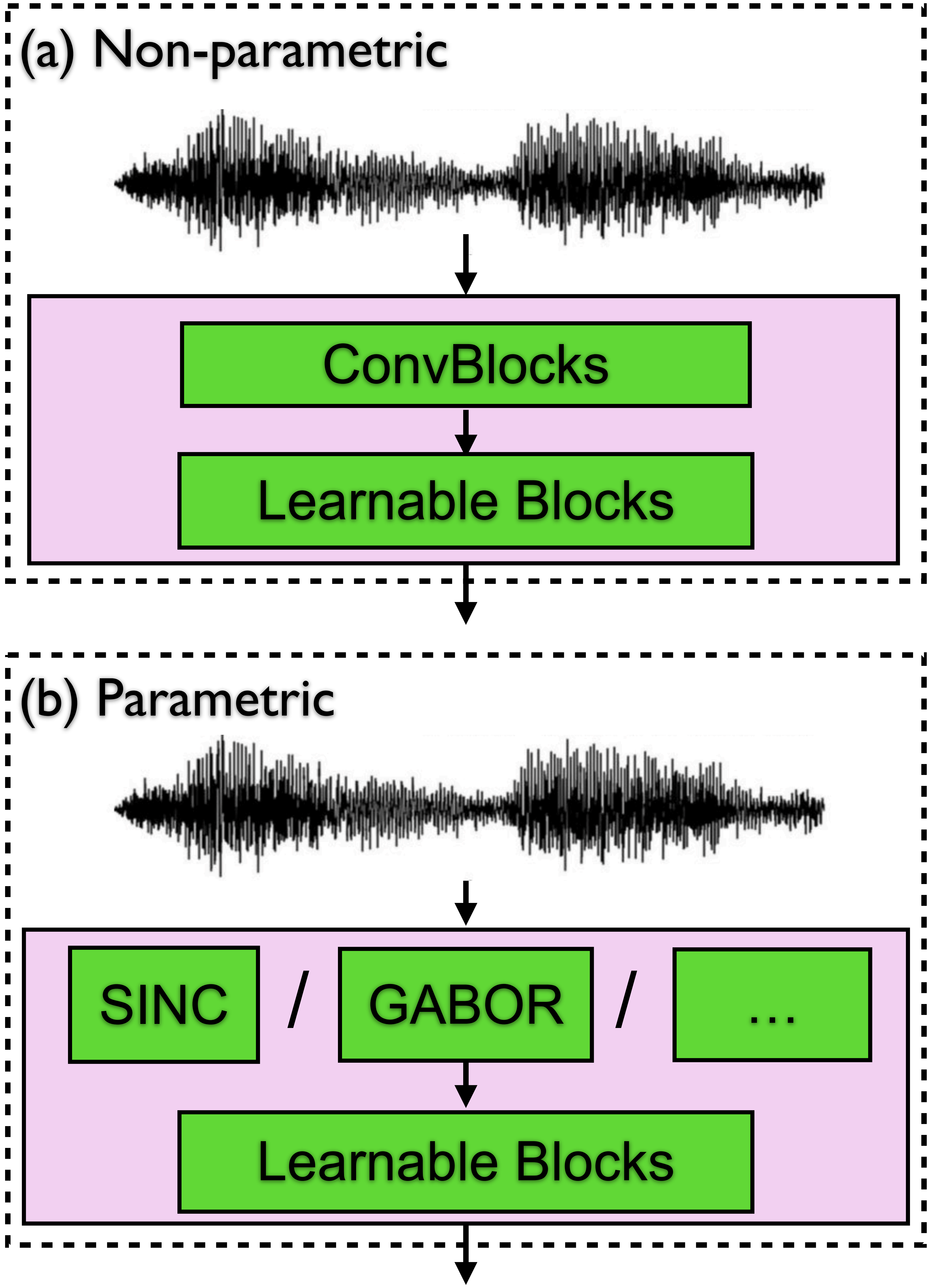
Model	Fix pre-train	Vox1-O	Vox1-E	Vox1-H
ECAPA-TDNN	-	0.87	1.12	2.12
HuBERT large	Yes	0.888	0.912	1.853
Wav2Vec2.0 (XLSR)	Yes	0.915	0.945	1.895
UniSpeech-SAT large	Yes	0.771	0.781	1.669
WavLM large	Yes	0.59	0.65	1.328
WavLM large	No	0.505	0.579	1.176
+Large Margin Finetune and Score Calibration				
HuBERT large	No	0.585	0.654	1.342
Wav2Vec2.0 (XLSR)	No	0.564	0.605	1.23
UniSpeech-SAT large	No	0.564	0.561	1.23
WavLM large (New)	No	**0.33**	**0.477**	**0.984**

Speaker verification

Table: GitHub repo for WavLM: Large-Scale Self-Supervised Pre-training for Full Stack Speech Processing

Prior Works





Left Figure: Waveform-based music processing with deep learning. ISMIR 2019 Tutorial

Channel Mismatch Problem

- Filters in the first layer conduct quasi time-frequency analysis, but tend to capture task-irrelevant aspects of the waveforms

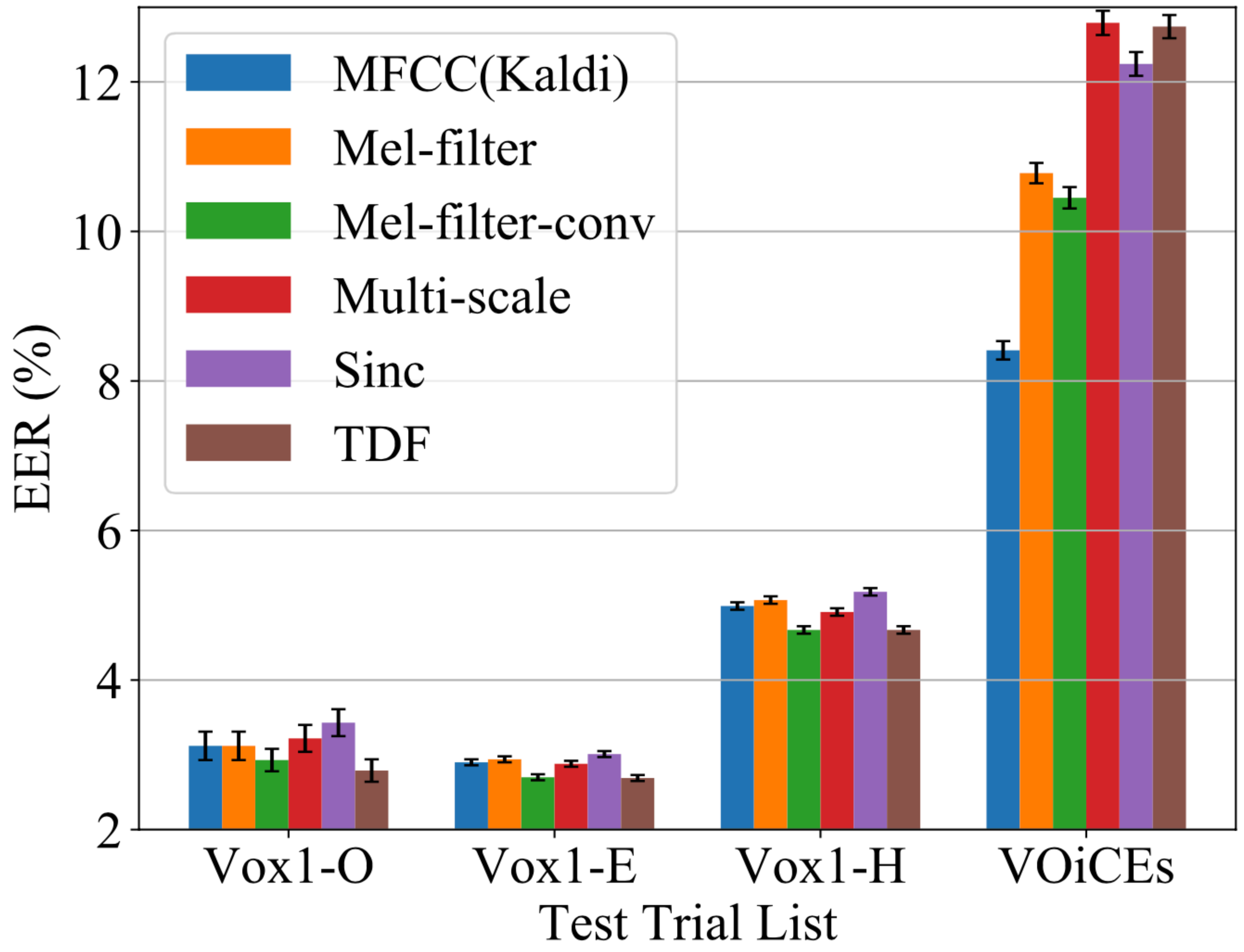
Different audio frontend SV performance under channel mismatch

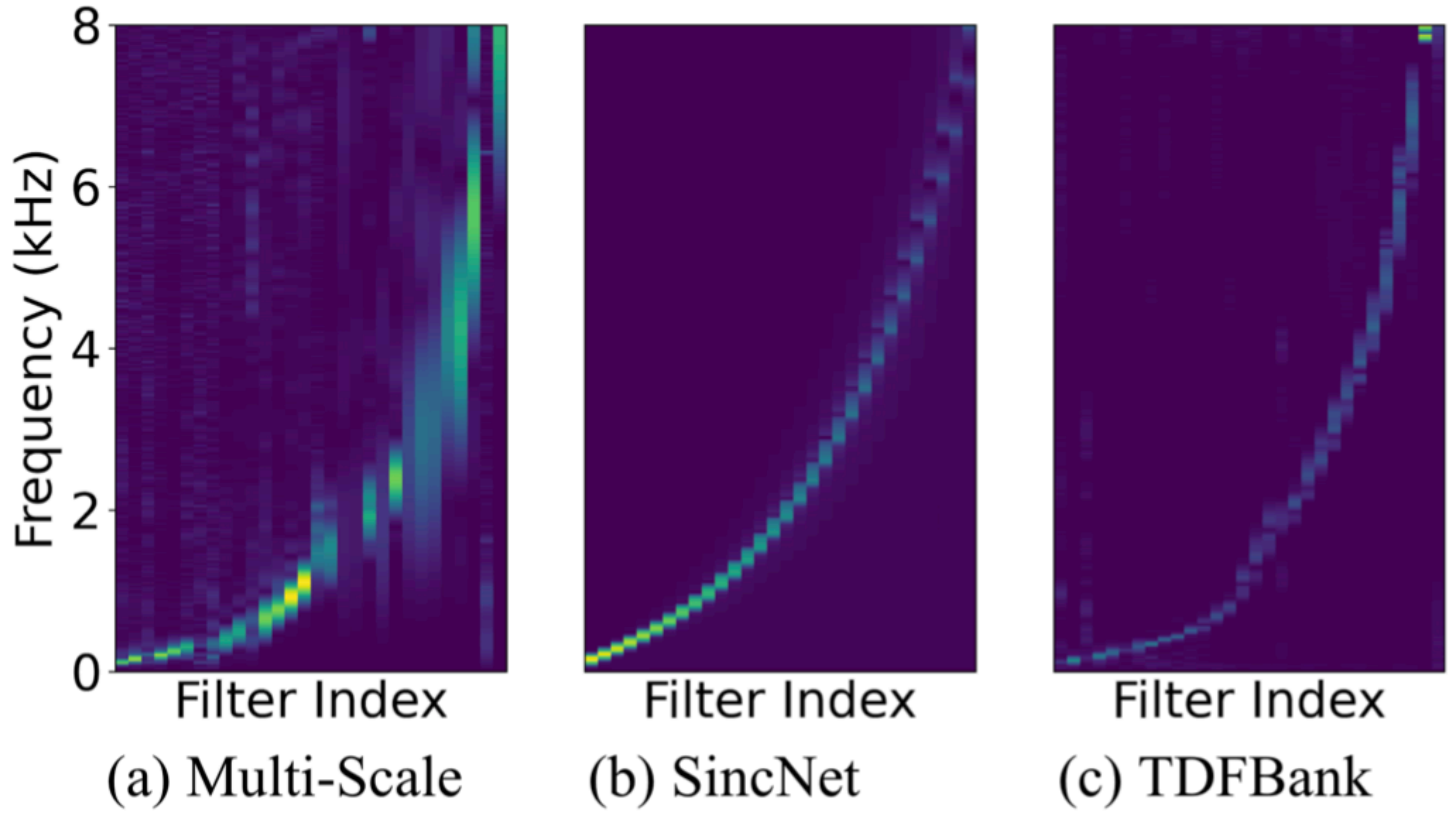
Experimental design:

- Train dataset: augmented VoxCeleb2
- Test dataset: Full VoxCeleb1 (in-domain) and VOiCEs (out-of-domain)
- Audio frontends: MFBank, Sinc, TDF, MultiScale with 25ms long, 30 channels/filters
- Common backbone for embedding network

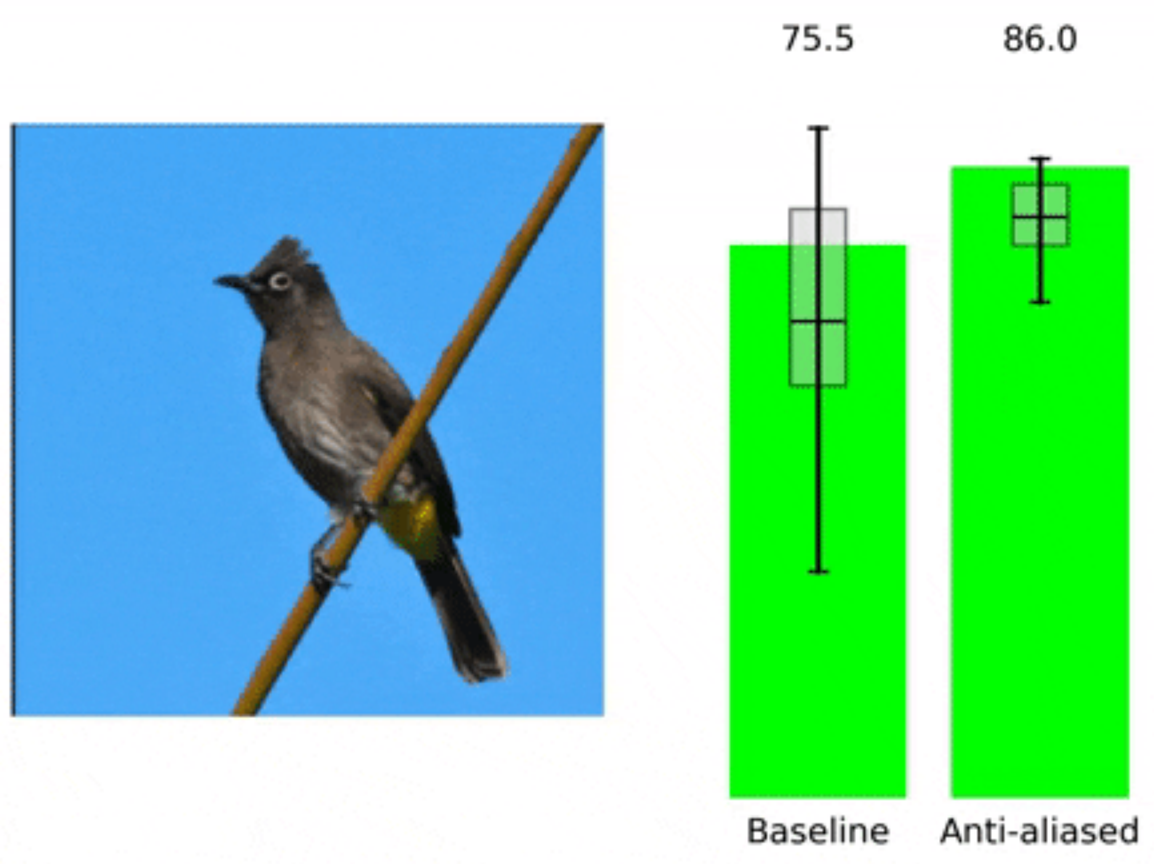
Different audio frontend SV performance under channel mismatch

- Results:





Proposed strategies



- Analytical Filters: modulus of filtered signal is shift-invariant

$$u_{\text{analytic}}(t) = u(t) + j\mathcal{H}[u(t)]$$

*Shift-invariant time-frequency representation

A general form of a magnitude-wise shift invariant linear time-frequency representation given signal x(t):

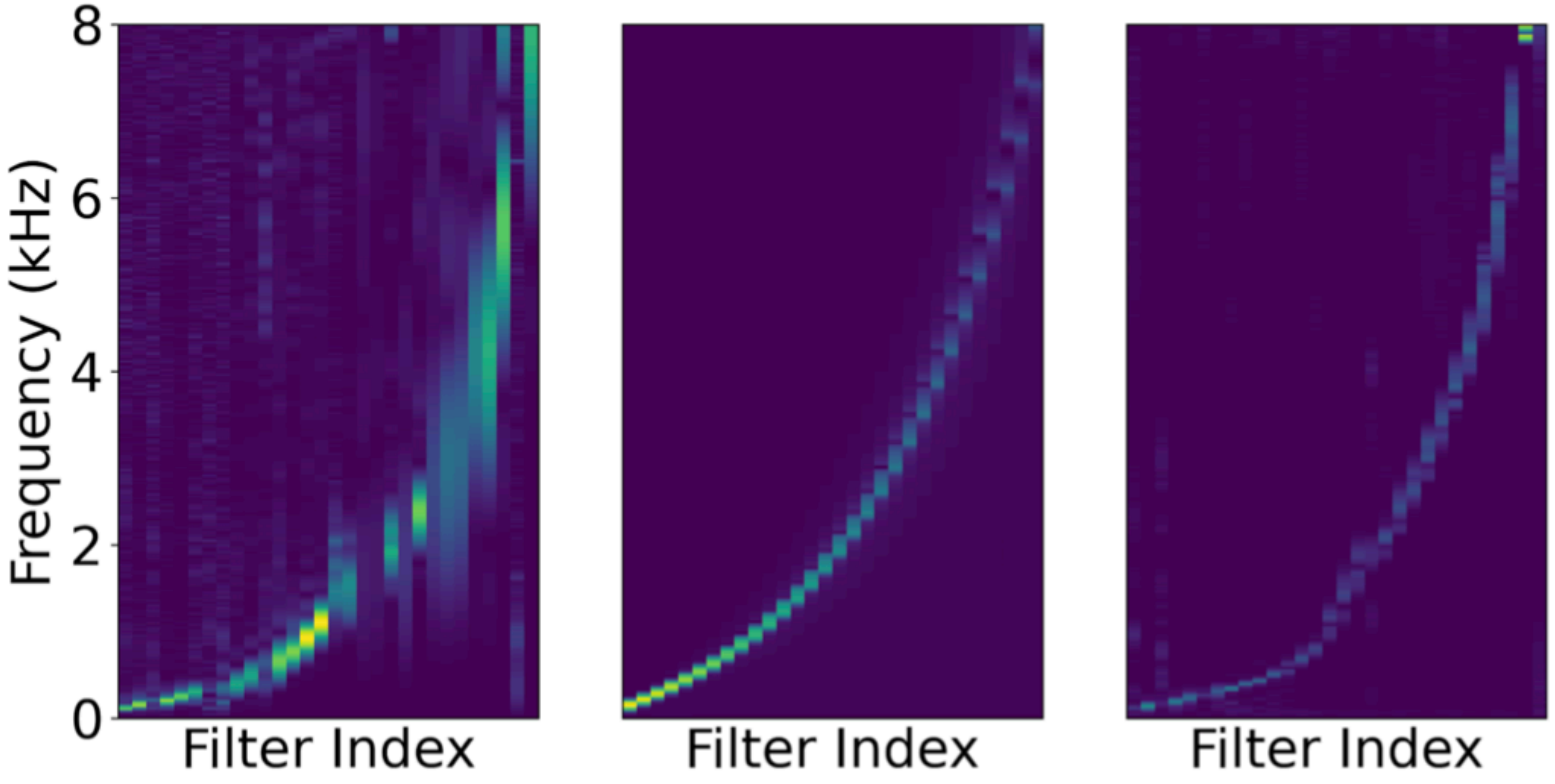
$$D_x(t, f) = \int g\left(t' - t\right) x\left(t'\right) e^{-j 2\pi f t'} dt'$$

Analytic representation assuming filterbanks are narrowband models:

$$\begin{aligned} u_a(t) &= u_m(t) \cdot \cos(\omega t + \phi) + i \cdot u_m(t) \cdot \sin(\omega t + \phi) \\ &= u_m(t) \cdot \left[\cos(\omega t + \phi) + i \cdot \sin(\omega t + \phi)\right] \\ &= u_m(t) \cdot e^{i(\omega t + \phi)}. \end{aligned}$$

Lütfiye Durak and Orhan Arıkan *Short-Time Fourier Transform: Two Fundamental Properties and an Optimal Implementation.* IEEE TRANSACTIONS ON SIGNAL PROCESSING
Hilbert transform. WIKIPedia

Proposed strategies



(a) Multi-Scale (b) SincNet (c) TDFBank

- Variational dropout on learned noisy filterbanks:

  Discard noisy filterbank weights in a smart way

Proposed strategies

Variational dropout

- Dropout: multiplying masks to NN weights.

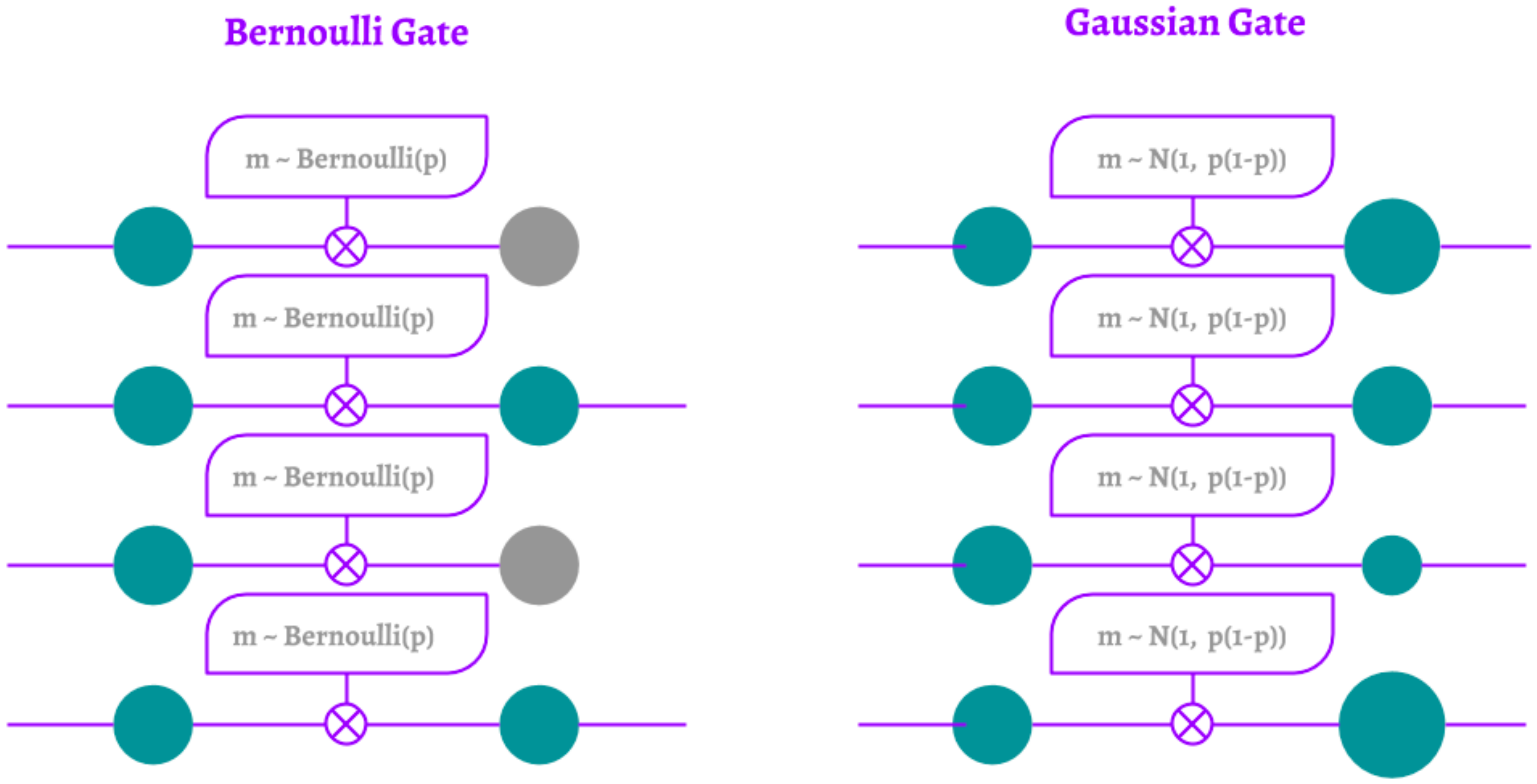


image: towardsdatascience.com/12-main-dropout-methods-mathematical-and-visual-explanation

Proposed strategies

Variational dropout

- Gaussian dropout: $\alpha = \dfrac{p}{1-p}$ is fixed

$$w_{ij} = \theta_{ij}\xi_{ij} = \theta_{ij}(1 + \sqrt{\alpha}\epsilon_{ij}) \qquad \epsilon_{ij} \sim \mathcal{N}(0, 1)$$

- Variational dropout: α_{ij} is learned for each weight

$$w_{ij} = \theta_{ij}(1 + \sqrt{\alpha_{ij}} \cdot \epsilon_{ij})$$

At inference:
$\alpha_{ij} >$ Threshold, drop the weights

Experiments

(1) Ablations on improvement of analyticity:

System	Vox1-O	Vox1-E	Vox1-H	Voices
x-vector (Kaldi)	3.12	2.9	4.99	8.41
x-vector	3.12	2.94	5.07	10.78
x-conv-vector	2.93	2.7	**4.67**	10.45
TDF	2.79	**2.69**	4.67	12.74
TDF+VD	3.01	2.79	4.81	11.10
TDF+$\mathcal{H}$	**2.72**	2.81	4.86	10.72
TDF+$\mathcal{H}$+BD	3.06	2.77	4.83	11.69
TDF+$\mathcal{H}$+GD	2.98	2.73	4.83	11.29
TDF+$\mathcal{H}$+VD	**2.72**	2.72	4.72	**10.32**

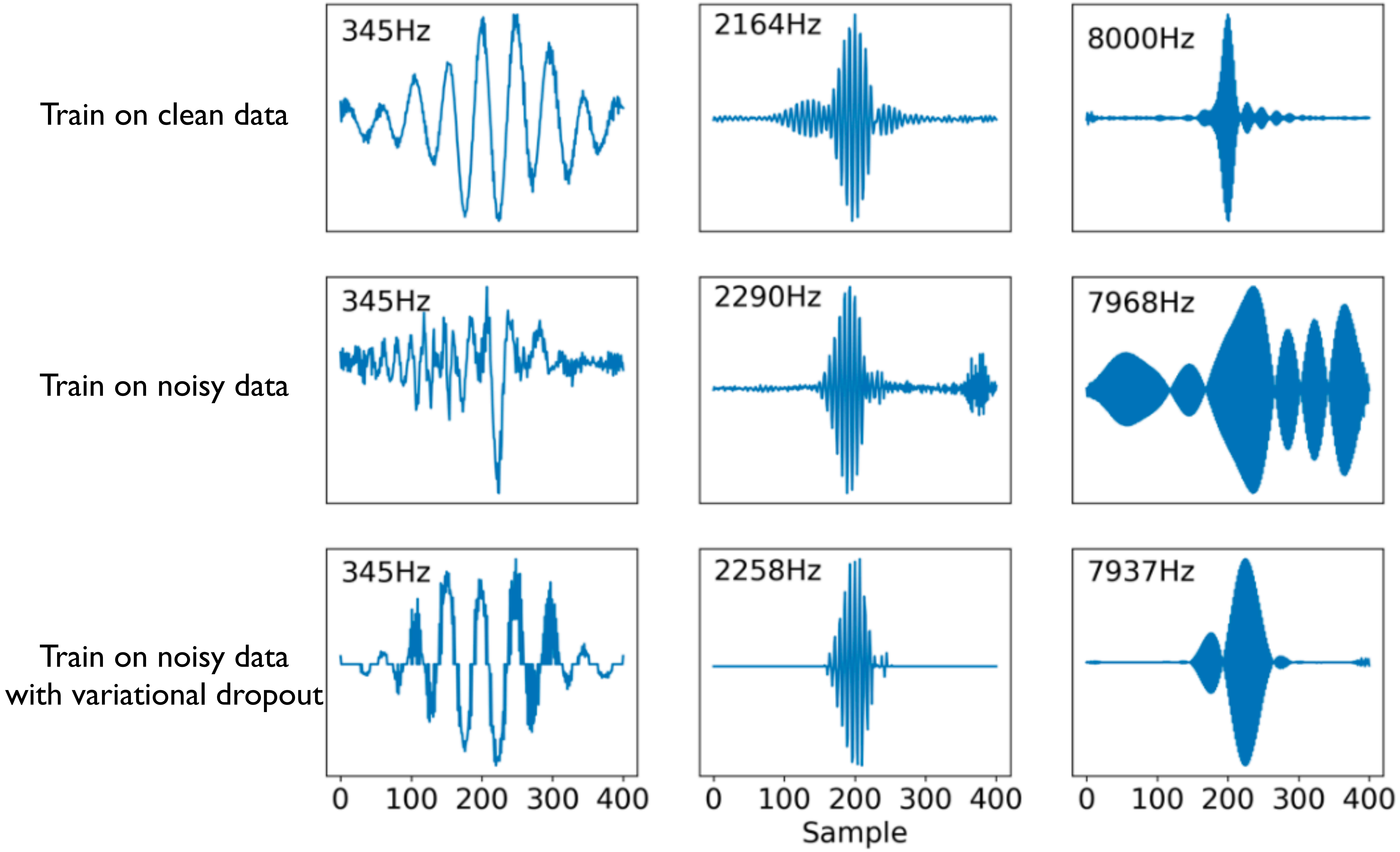
Experiments

(2) Ablations on improvement of variational dropout:

System	Vox1-O	Vox1-E	Vox1-H	Voices
x-vector (Kaldi)	3.12	2.9	4.99	8.41
x-vector	3.12	2.94	5.07	10.78
x-conv-vector	2.93	2.7	**4.67**	10.45
TDF	2.79	**2.69**	4.67	12.74
TDF+VD	3.01	2.79	4.81	11.10
TDF+$\mathcal{H}$	**2.72**	2.81	4.86	10.72
TDF+$\mathcal{H}$+BD	3.06	2.77	4.83	11.69
TDF+$\mathcal{H}$+GD	2.98	2.73	4.83	11.29
TDF+$\mathcal{H}$+VD	**2.72**	2.72	4.72	**10.32**

Experiments

(2) Variational dropout



Channel mismatch in waveform speaker embedding modeling

Experiments

(2) System comparisons

System	Feature	VoxCeleb-O		VoxCeleb-E		VoxCeleb-H		VOiCEs	
		EER	min-DCF	EER	min-DCF	EER	min-DCF	EER	min-DCF
x-vector (Kaldi)	MFCC	2.26	0.256	2.37	0.279	4.14	0.408	**6.79**	**0.553**
x-vector	Mel-fbank	2.37	0.264	2.42	0.280	4.18	0.406	8.14	0.658
x-conv-vector	Mel-fbank	**2.04**	**0.241**	**2.17**	**0.252**	**3.79**	**0.379**	7.10	0.581
Multi-scale	Waveform	2.28	0.273	2.38	0.285	4.17	0.408	8.54	0.705
Sinc	Waveform	2.37	0.287	2.32	0.278	4.02	0.400	8.55	0.682
Sinc+$\mathcal{H}$	Waveform	2.15	0.270	2.28	0.271	3.91	0.396	8.90	0.669
TDF	Waveform	**1.98**	**0.230**	**2.19**	**0.249**	**3.85**	**0.383**	8.38	0.663
TDF+$\mathcal{H}$	Waveform	2.01	0.261	2.27	0.263	3.98	0.396	7.46	**0.621**
TDF+VD	Waveform	1.98	0.235	2.30	0.264	4.05	0.385	7.68	0.626
TDF+$\mathcal{H}$+VD	Waveform	1.99	0.266	2.26	0.253	3.93	0.385	**7.40**	0.633

Experiments

(2) System comparisons

System	Feature	VoxCeleb-O		VoxCeleb-E		VoxCeleb-H		VOiCEs	
		EER	min-DCF	EER	min-DCF	EER	min-DCF	EER	min-DCF
x-vector (Kaldi)	MFCC	2.26	0.256	2.37	0.279	4.14	0.408	**6.79**	**0.553**
x-vector	Mel-fbank	2.37	0.264	2.42	0.280	4.18	0.406	8.14	0.658
x-conv-vector	Mel-fbank	**2.04**	**0.241**	**2.17**	**0.252**	**3.79**	**0.379**	7.10	0.581
Multi-scale	Waveform	2.28	0.273	2.38	0.285	4.17	0.408	8.54	0.705
Sinc		2.37	0.287	2.32	0.278	4.02	0.400	8.55	0.682
Sinc+$\mathcal{H}$		2.15	0.270	2.28	0.271	3.91	0.396	8.90	0.669
TDF		**1.98**	**0.230**	**2.19**	**0.249**	**3.85**	**0.383**	8.38	0.663
TDF+$\mathcal{H}$		2.01	0.261	2.27	0.263	3.98	0.396	7.46	**0.621**
TDF+VD		1.98	0.235	2.30	0.264	4.05	0.385	7.68	0.626
TDF+$\mathcal{H}$+VD		1.99	0.266	2.26	0.253	3.93	0.385	**7.40**	0.633

Conclusions

- We studied cross channel speaker verification performance of raw-waveform based speaker embeddings

- We proposed to introduce (1) analyticity and (2) variational dropout to alleviate the performance mismatch